The problems in a Question Answering system in the academic domain

P.López-Moreno, A.Ferrández, S.Roger, S.Ferrández
Natural Language Processing and Information Systems Group
Department of Software and Computing Systems
University of Alicante, Spain
{P.Lopez}@ua.es
{antonio,sferrandez,sroger}@dlsi.es

Abstract

In this paper we present BRUPIA, a Question Answering (QA) system in a restricted domain: the web academic environment at the University of Alicante. This system is the transformation of an open domain QA system, AliQAn. This paper focuses on explaining how an open domain system can be transformed into another one that can successfully work on a web restricted domain. We analyze the problems of carrying out this task and we also develop the necessary resources for the new system like the corpora, the questions and the set of patterns in the new domain. Finally, a new strategic approach for the improvements in the use of the terminology in web domain is proposed. The measure of evaluation is the Mean Reciprocal Rank and the final result is 32,5%.

Keywords

Question answering system academic restricted domain web

1 Introduction

Different trends are used in Question Answering (QA) systems. On the one hand, the traditional QA focuses to process large amount of independent documents. Some of these systems takes part in the TREC and CLEF evaluation campaigns. On the other hand, there are systems based on the web. Most of these systems represent the restricted domain, and they use the websites to extract documents that have fixed structures and are related by means of a hierarchy of pages. The most important difference between these QA systems, is that the first one works with independent documents in a journalist style and the second one has a document collection with a web structure. The information appearing in these websites can be distributed in different texts, even in different related documents just by means of using links. It is very difficult to find the textual information as this one appears in the question. We will take an example of the queries like: *Quién es el director del DLSI? (Who is the manager of the DLSI?)*. In this case, the correct answer appears in the organization site of the Department of Software and Computing Systems (DLSI), but the sequence "manager of DLSI" is not in the text, DLSI appears at the top of the page. The information of the page is referred to the staff working at this department, so that, the head contains the word "DLSI" and the content gives details of the different positions among them, like the manager, enclosed to their names. Another example is the following: *Qué página personal tiene Antonio Ferrández? (What is the personal website of Antonio Ferrandez?)*. Looking for the answer, we must go to the main page of DLSI department and click on "Teaching Staff". A list of teachers is shown, where each name is a link to its personal page. But in this situation, the distance between the name and the correct answer has a main role. These examples are typical cases in a web domain.

Analyzing the restricted domain QA systems, we can make out two different tendencies. First, a restricted domain QA system has a baseline for open domain as a point of starting. This way, a general system can be transformed into a specific one. Furthermore, a QA system can be created directly for the specific domain.

We develop the first technique for our QA system in the academic domain, considering the AliQAn system as a starting point. Moreover, the specific terminology of this domain and the web structure are used in order to improve the accuracy of our system. The results obtained with the measure of evaluation Mean Reciprocal Rank (MRR) are 32,5%.

The rest of the paper is structured as follows: First, we introduce the backgrounds in a restricted domain system. Secondly, we summarize the characteristics of the restricted domain systems based on Web. Next, we explain the process to transform the baseline in the new system BRUPIA. Finally, we show the problems and the solutions found, and the main conclusions obtained.

2 Backgrounds

Many QA systems employed sources in order to store the specific terminology. For instance ExtrAns [5] is a QA system aimed at restricted domains, in particular terminology-rich domains. They carried out "terminological normalization", where a term is replaced by a synset identifier when this term belongs to the category in the terminology knowledge represented by means of an ontology. The document collections in the

genomic domain was generated from Medline. In this way, although the system uses the web to extract the documents, these texts are independent.

Another example is a QA system for a home agent robot [1], which is based on templates to store the data. Each expected question topic is defined as a single query frame and each frame has a rule for SQL generation. The web crawler downloads the selected webpages from the website of the Korea Meteorological Administration and the wrapper is used to extract weather information from the webpages stored in a database.

The next study case is a system developed for the company Bell Canada to answer to client's questions in services offered by a big company [3]. They experimented with some methods of reranking with information about the domain specific language, particularly with vocabulary issues. In this case, the document collection was derived from .html and .pdf files as our system. As the structure of these files was so complicated, documents were saved as pure texts sacrificing some elements like titles, listings or tables.

In general, each restricted domain QA system uses some different techniques to the treatment of the terminology, because this one is the main role in this kind of system. In addition to this fact, there are systems that are based on an initial baseline. Some systems are based on the web and download directly the documents, while others have a document collection with independent texts.

Our approach combines some of these techniques. First, BRUPIA has a baseline system in open domain. Secondly, our system uses the web to obtain the documents and finally, we use the specific terminology of the domain in a strategic way.

3 Characteristics of restricted domain

The first section describes general characteristics based on [1, 2, 3].

1. Quality of responses must be higher because of the practice on the market.
2. The answers are searched in relatively small domain collections, so the redundancy is lower than in an open domain system.
3. User requirements in the quality of the answer tend to be higher in restricted domains. No answer is preferred to a wrong answer.
4. The terminology plays a central role.

Our particular contributions about web domain are represented in the following paragraphs:

1. The structure of the web documents lets to identify a symbolic structure, so it is possible to split different parts of the document in order to provide a higher score and to obtain better results.
2. Webpages contain dependent information and have a hierarchy of pages. The related information with one question can be separated in some documents, or it even can be necessary to visit different pages to find the correct answer.

4 Transformation of the AliQAn into BRUPIA

We propose a monolingual Spanish QA system named BRUPIA for an academic domain, particularly, the domain of the University of Alicante (UA). From our baseline AliQAn, a monolingual open domain QA system developed at the UA three years ago, we have adjusted the new system modifying the patterns and applying the necessary techniques to the treatment of the new domain knowledge. AliQAn participated in the CLEF-2005 [6] competition and last year, it participated in the CLEF-2006 [7] with a new version of our system.

We have had some problems with the new system BRUPIA. After that, we will explain a detailed description about the problems detected and the solutions proposed.

There are three large groups of problems. First of all, the generation of the corpus. We experimented with two different collections in the academic domain. Both collections were generated automatically from pages of the UA. The size of the first corpus is 102.900 documents. The second collection is more concrete than the first one. It was constituted by documents of the web but considering only the domain of the DLSI. Finally, this corpus contains 2.900 documents.

The second kind of problem is related to the system questions about an insufficient typology and several difficulties to allocate the correct type of the questions.

Finally, we analyze some problems in the baseline system regarding to the patterns.

4.1 Problems in the generation of the corpus

4.1.1 IRn problems

IRn is a passage retrieval that returns a list of relevant documents for each question. The web structure causes that IRn returns the document in an incorrect way. For example the question: *Qué es DLSI? (What is DLSI?)*, some documents of the corpus contain the correct answer for this question, where the word "DLSI" appears with its description, but the occurrence of this word is very low, so these documents are not returned by IRn. Nevertheless, there are documents that contain sometimes the word "DLSI", such as, the page of the staff of this department, where this word appears in the email of each person. These documents appear in the returned list by IRn but BRUPIA system cannot find the solution.

4.1.2 Parsing problems

Another problem derived from the web structure is the malformation of the syntactic blocks (SB). The new documents are very different from the initial documents of the baseline system, as for as the sentences segmentation and the style of the texts. The main problem is the lack of the full-stop or period to indicate the final of the sentences. So, our parser SUPAR carries out the wrong formation of the SB because it is not able to separate correctly the different blocks. The following Figure 1 shows the malformation of the

SB caused by the absence of a point at the end of the sentences due to the automatic conversion of the text.
Secretario: Juan Antonio Pérez Ortiz (Secretary: Juan Antonio Perez Ortiz)
Subdirector: Patricio Martínez Barco (Assistant principal: Patricio Martinez Barco)

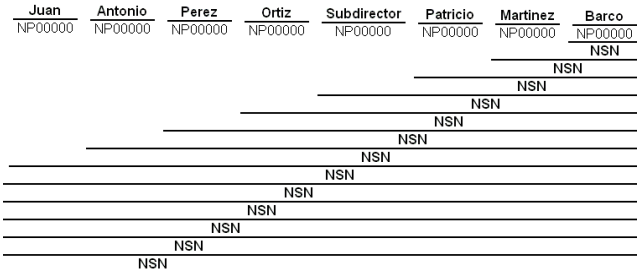

Fig. 1: *Example of the segmentation of SUPAR*

4.1.3 Problems caused by the web structure

The information in a HTML document is semi-structured and it can be related with other pages. The style in this kind of documents is totally different from other any style. At the top, there is a page head or a title which summarizes the content of the document. If we select the webpage about the staff of this department, we can see the information represented in a table format with the names of the members of the organization. There is a question like: *Quién es el secretario del DLSI? (Who is the secretary of the DLSI?)*. In this document, the word secretary and the name appear, but the literal "DLSI" is not there, because it appears in the head.

Another problem is produced when the distance between the question words and the solution is very long. For instance, if the system looks for the projects that one teacher is carrying out: *En qué proyectos participa Patricio? (What projects does Patricio work in?)*. The name of the teacher appears at the top of the page, and then, all information appears like a list with different items, and the projects are at the end of this list. All information in the page is about the same teacher, but the name only appears in the head. So, when we look for the information there is too big distance and the score is too low.

4.1.4 Problems because of the languages

This system is monolingual, so the language of the documents must be the same that the questions. These documents were downloaded automatically, so, there are some documents in different language. Sometimes, the URL indicates the language that and the documents can be leaked. However, others do not contain any indicators of the language, and when the system returns the answer is different from the question language.

4.2 Problems detected in questions

4.2.1 Problems of allocation of correct type

The system BRUPIA has a collection of the 100 questions. In accordance with the baseline typology, some questions were classified with an incorrect type. The lack of information in the question was the first reason, for example: *Cuál es el número de la centralita de la Universidad? (Which is the number of the switchboard of the University?)*. We know that this number is referred to the phone number but the system interprets the type as a quantity.

Besides, the classification patterns do not contain all the options, one example of this situation is: *Qué dirección electrónica tiene Loren Moreno Monteagudo? (What electronic direction does Loren Moreno Monteagudo have?)*. The system determines incorrectly that the type is group instead of the email type.

There are wrong cases with specific concepts for this domain, like the word "extension", which is used for the telephone line inside of the academic domain, but as measure in the baseline system. For this situation, there is a question like: *Qué extensión tiene Jesús Peral Cortés? (What extension number does Jesus Peral Cortes have?)*.

4.2.2 Insufficient typology

Initially, the baseline had a typology with the following concepts: profession, first name, person, group, place country, place city, capital place, place, abbreviation, event, object, weather date, weather year, weather month, weather day, weather events, numerical economic, numerical quantity, numerical percentage, numerical measurement, numerical period, numerical age, definition, email, telephone and fax. This classification is scarce for the new domain and the system needs more concepts.

4.3 Problems in the baseline

4.3.1 Problems with the patterns

The major problems are the definitions. The journalistic style of the corpus of the baseline is composed by narrative texts, therefore the definition is more probably that appears before the term. The new domain has a different style, so in some definitions, the concept appears in the first place and afterwards the definition. So, it is more interesting to look up the definition on the right and the term on the left, modifying some parameters of the patterns.

5 Representation of problems

The seven types of errors detected in the adaptation to new academic domain are represented in the Figure 2. The types of problems and their percentages are represented in the graph. With regard to the colours, each type of error has one different colour, but there are two special situations like the parsing problems and the problems because of incorrect allocation of the type, which have combined colour to indicate that these failures are due to other more general problems.

	Insufficient typology	Problems in patterns	Problems of IRn	Parsing problems	Problems because of the web structure	Problems with languages	Allocations of the incorrect type
Number of questions	12	2	7	13	10	4	12
Error percentage	20 %	3 %	12 %	22 %	17 %	7 %	19 %

Table 1: *Representations of problems*

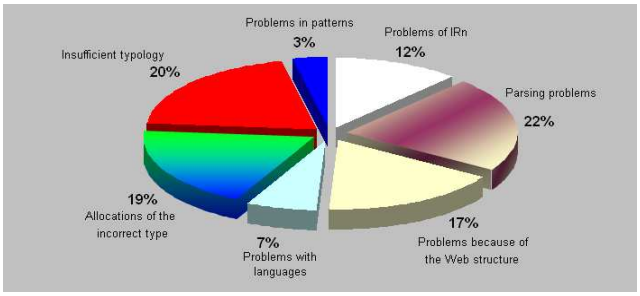

Fig. 2: *Representation of problems*

Concretely, the parsing problems are due to the web structure and the allocation of incorrect type is because of failures of the patterns. Most important problem are caused by the web structure that generates an error percentage of 39% *(22% parsing problems caused by the web structure and 17% problems because of web structure)*.

In the Table 1, it is possible to distinguish the number of the questions and the error percentage of each type of error.

6 Proposed solutions

Once the errors are detected, we show viable solutions for the different failures presented in the previous section.

6.1 Solution for IRn and SUPAR

The reason of these problems consists in how the web is structured, concretely the lack of punctuation marks to indicate the end of the sentences. Along these lines, we propose to add a point at the end of the sentences to solve the segmentation of these sentences. In this way, we could solve the problem of malformation of SB carried out for SUPAR, and the problems of IRn, both caused by an incorrect segmentation of the sentences.

6.2 A method to improve the precision

In this point, our best innovation is presented. It consists in applying a method to use the information that pertains to different hierarchy levels in a strategic way. The related data are distributed in different texts or at least in different places of the document. Usually, the main information appears in the page head or in the title of the document. Words appearing in the head usually are not more times in the text, because these terms are general and describe all the information that is contained in the webpage. So, we use the most important terminology that appears in a web document. We look for the words of the question which appears in the page head of the document and we remove them to the question to not look for them in the document. At the same time, we keep the definition of these words that are removed to question when the system detects that the definition question is referred to these terms. For example, for the question introduced in the initial part of this document: *Quién es el director del DLSI? (Who is the manager of the DLSI?)*. The text "DLSI" only appears in the head. To solve this question, the system removes the word "DLSI" when it detects this word in the question, and it only looks for the text "director (manager)" in the document, returning the name of the person.
Another kind of failure takes place when the information is separated by means of a high distance. For instance, a question mentioned previously: *En qué proyectos participa Patricio? (What projects does Patricio work in?)*. In the document, this name appears at the top, however the projects appear at the end of the page. In this case, we propose to remove the word "Patricio" to the question and to look for only the "proyectos (projects)", solving the problem of the distance.

6.3 Solution for language problems

Two points of view are possible to solve these problems. On one hand, we could filter the language of the documents and create different corpus for each language and use the spanish corpus for the monolingual system. In the future, the other ones will be treated. So, we want to use a resource developed in this department [4], which will allow us to detect the language of the documents and to leak the spanish ones.

6.4 Solution for problems with the patterns

In order to solve the problems of incorrect allocation of types, the classification patterns were adapted to the new group of questions, extending the conditions.

Besides, the extraction patterns were adapted to improve the precision in our system. So that, the definition is looked after to the acronym because of the probability of appearing with this format is greater.

6.5 Solution for insufficient typology

The typology must be extended considering other types like: personal page, guardianship schedule, office, subject, course and mailing dress. In addition,

1st answer	2nd answer	3rd answer
27	9	3

Table 2: *Results of BRUPIA*

some types were adapted for the new system BRUPIA.

7 Representation of solutions

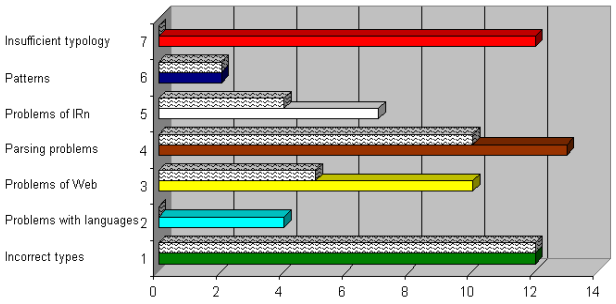

Fig. 3: *Problems and solutions*

Two important things are represented in the Figure 3. Firstly, the detected errors that are represented in the line of down and marked with the concrete colour of each type of error and secondly, the solved errors that are situated over the detected ones. For each type, it is possible to check the obtained improvement comparing the size of both lines. The different errors detected are represented in the "y" axis, and the number of questions that have each problem is detailed in the "x" axis. The errors of classification and extraction patterns have been solved totally. Besides, some errors of the web structure have been clarified with our special contributions, ignoring the proper names of the questions that are contained in the page head, which pertain a higher hierarchic level. However, the used typology is the same than the baseline and the failures related to this concept have not been solved yet. In this way, the problem of language is future work. Even so, the final result is 32,5% of MRR.

8 Results

Regarding the first experiment carried out with the general training corpus for UA, the obtained results were about 5% the precision. The final result considering a set of 100 questions for the restricted domain of the DLSI is 32,5% of MRR.

In the Table 2, it is possible to distinguish the number of questions that are correct in the first, second or third position. Moreover, two answers in Valencian language returned in the second position have been considered correct. It is very interesting that the number of correct answers returned in first position is higher than the other groups.

Acknowledgments

This research has been partially funded by the Spanish Government under project CICyT number TIN2006-15265-C06-01 and by the University of Comahue under the project 04/E062.

This work has been partially supported by the EU funded project QALL-ME (FP6 IST-033860).

9 Conclusion and future work

BRUPIA is a QA system for academic domain of UA. Our approach is different from the traditional QA systems, which works with independent documents. BRUPIA has a document collection with structured information and contains related data. In addition, these documents have a hierarchy of webpages that indicates how the texts are related. Our system uses the terminology of the domain in a strategic way to solve some problems with the web structure. The patterns have been adapted to new domain generalizing the conditions. Finally, we propose some alternatives to solve the specific problems in this kind of web systems. The results obtained with our experiments are 32,5% of MRR, obtaining important improvements with respect to the initial tests.

In the earliest phases, BRUPIA solved some of the problems that came up in the adaptation of the baseline system to the new domain. Nevertheless, much work is left for our future work to generate a robust system for the academic domain of UA.

References

[1] H. Chung, Y.-I. Song, K.-S. Han, S.-H. Kim, D.-S. Yoon, J.-Y. Lee, and H.-C. Rim. A practical QA System in Restricted Domains. *In Proceedings of the ACL Workshop*, pages 39–45, July 2004.

[2] D. Ferrés and H. Rodríguez. Experiments Adapting an Open-Domain Question Answering System to the Geographical Domain Using Scope-Based Resources. *In Proceedings of the Multilingual Question Answering Workshop of the EACL*, pages 69–76, April 2006.

[3] D.-N. Hai and K. Kosseim. The problem of Precision in Restricted-Domain Question-Answering.Some Proposed Methods of Improvement. *In Proceedings of the ACL Workshop*, pages 8–15, July 2004.

[4] T. Martínez, E. Noguera, R. Muoz, and F. Llopis. Web track for CLEF2005 at ALICANTE UNIVERSITY. *In Workshop of Cross-Language Evaluation Forum (CLEF)*, September 2005.

[5] F. Rinaldi, J. Dowdall, G. Schneider, and A. Persidis. Answering Questions in the Genomics Domain. *In Proceedings of the ACL Workshop*, pages 46–53, July 2004.

[6] S. Roger, S. Ferrández, A. Ferrández, J. Peral, F. Llopis, A. Aguilar, and D. Tomás. AliQAn, Spanish QA System at CLEF-2005. *In Workshop of Cross-Language Evaluation Forum (CLEF)*, 2005.

[7] S.Ferrández, P.López-Moreno, S.Roger, A.Ferrández, J.Peral, X.Alvarado, E.Noguera, and F.Llopis. AliQAn and BRILI QA systems at CLEF 2006. *In Workshop of Cross-Language Evaluation Forum (CLEF)*, 2006.